Under-determined source separation: comparison of two approaches based on sparse decompositions

Sylvain Lesage, Sacha Krstulović and Rémi Gribonval
firstname.lastname@irisa.fr

METISS project, IRISA-INRIA
Campus de Beaulieu, 35042 Rennes Cedex, France

Abstract This paper focuses on under-determined source separation when the mixing parameters are known. The approach is based on a sparse decomposition of the mixture. In the proposed method, the mixture is decomposed with Matching Pursuit by introducing a new class of multi-channel dictionaries, where the atoms are given by a spatial direction and a waveform. The knowledge of the mixing matrix is directly integrated in the decomposition. Compared to the separation by multi-channel Matching Pursuit followed by a clustering, the new algorithm introduces less artifacts whereas the level of residual interferences is about the same. These two methods are compared to Bofill & Zibulevsky's separation algorithm and DUET method. We also study the effect of smoothing the decompositions and the importance of the quality of the estimation of the mixing matrix.

1 Introduction

The source separation problem [1] consists in retrieving unknown signals (the sources) from the only knowledge of mixtures of these signals (the channels). Each channel x_n is the sum of the filtered sources :

$$x_n(t) = \sum_{i=1}^{I} (a_{n,i} * s_i)(t) \tag{1}$$

where $a_{n,i}$ are filters. When the mixture is linear instantaneous, the filters correspond to a multiplication by a constant. Thus the mixture can be written in linear algebra as $\mathbf{x} = \mathbf{A}\mathbf{s}$, where $\mathbf{A}$ is the mixing matrix, and the rows of the matrices $\mathbf{x}$ and $\mathbf{s}$ are respectively the signals x_n and s_i. In the determined (resp. over-determined) case, where the number of observed channels is equal to (resp. greater than) the number of sources, estimating the mixing matrix and estimating the sources are equivalent problems. Conversely, in the under-determined case, the knowledge of the mixing matrix or its estimate is not sufficient to recover the sources, and a model of the sources is generally needed to estimate them [2]. Generally, it is a difficult task to distinguish, in the performances of a given algorithm, the effect of the quality of the matrix estimation from the effect of the mismatch to the model.

In this article, we focus on the under-determined case. Our approach uses models based on the existence of sparse representations of the sources [3], and assumes the perfect knowledge of the mixing matrix. We compare two separation algorithms based on variants of Matching Pursuit (MP) [4]. The first variant consists in decomposing the

multi-channel mixture without knowing the mixing matrix, and then using the mixing matrix to classify the coefficients of the decomposition and affecting them to the sources to estimate [5,6]. The second variant consists in using the mixing matrix in the sparse decomposition step itself, and no additional classification step is needed. The performance of these two algorithms are compared to the best linear separator (BLS) [7], to the Bofill & Zibulevski's algorithm (BZ) [8] and to the DUET algorithm [6].

This article is organized as follows : in section 2, we recall the general definition of Matching Pursuit. Multi-channel MP and its various separation algorithms are described in section 3 and we detail the experimental conditions and the results in section 4.

2 Matching Pursuit

A signal x (considered as a vector of the Hilbert space $\mathcal{H}$ of finite-energy signals) admits a sparse decomposition over the dictionary $\mathcal{D} = \{\phi_k\}$ of atoms ϕ_k – or elementary signals ϕ_k – if it can be written as a linear combination $x = \sum_k c_k \phi_k$ where few coefficients $\{c_k\}$ are non-negligible. In this framework, MP iteratively computes sparse approximations of the form $x = \sum_{m=1}^{M} c_{k_m} \phi_{k_m} + R^M$ where R^M is a residual that tends to zero as the number of iterations M tends to infinity. The principle of the algorithm is to select, at each step, the atom that is the most correlated to the residual, then to update the residual by removing the contribution of this atom.

The most current stopping criteria are based on the absolute or relative level of energy of the residual or/and on a fixed number of iterations to run. In addition, the Gabor dictionary is classically used to sparsely decompose audio signals. It is composed of a collection of time-frequency Gabor atoms $\phi_{s,u,\xi}(t) = w\left(\frac{t-u}{s}\right) \cdot \exp\left(2j\pi\xi(t-u)\right)$. These atoms are defined by the choice of a window w of unit energy (Hanning, Gaussian, ...), a scale factor s, a time localization u, and a frequency ξ. Such a dictionary allows a fast computation of the inner products between the signal and the atoms by applying some windowed-FFTs.

3 Source separation with Matching Pursuit

Source separation techniques based on sparse approximations of multi-channel signals on a dictionary have been proposed in the multi-channel case [3,6]. More specifically, in the MP framework, the method proposed in [5,9] uses multi-channel MP, followed by a clustering (note that the base idea of this method could be developed for other multi-channel sparse decomposition algorithms, e.g. [10].) After recalling the principle of the method based on MP plus clustering, we propose a variant where the definition of the dictionary includes knowledge of the mixing matrix $\mathbf{A}$.

3.1 Multi-channel Matching Pursuit

For the sparse decomposition of multi-channel signals, we use a dictionary $\boldsymbol{\mathcal{D}}$ composed of multi-channel atoms $\boldsymbol{\phi}$. These atoms are defined by $\boldsymbol{\phi} = (c_1\phi, c_2\phi, \ldots, c_N\phi)$, where $\phi \in \mathcal{D}$ is a mono-channel atom from a dictionary $\mathcal{D}$ and where the coefficients $c_1, \ldots, c_N$ satisfy $\sum_{n=1}^{N} c_n^2 = 1$. After M iterations, multi-channel MP leads to a decomposition of the form :

$$(x_1, \ldots, x_N) = \widehat{\mathbf{x}}^M + (R_1^M, \ldots, R_N^M).$$

with $\widehat{\mathbf{x}}^M := \sum_{m=1}^{M} (c_{1,k_m}\phi_{k_m}, \ldots, c_{N,k_m}\phi_{k_m})$. The algorithm is composed of the following steps :

1. Initialization : $M = 1, \quad R_n^0 = x_n, \quad c_{n,k} = 0, \quad \forall n, \forall k$;
2. Computation of the inner product between each channel of the residual R_n^{M-1} and each atom ϕ_k of the mono-channel dictionary.
3. Selection of $k_M = \arg\max_k \sum_{n=1}^{N} |\langle R_n^{M-1}, \phi_k \rangle|^2$
4. For each channel n, update of the residual : $R_n^M = R_n^{M-1} - \langle R_n^{M-1}, \phi_{k_M} \rangle \phi_{k_M}$ and of the coefficients : $c_{n,k_M}^M = c_{n,k_M}^{M-1} + \langle R_n^{M-1}, \phi_{k_M} \rangle$
5. If the stopping criterion has not been reached, $M \leftarrow M + 1$, then go back to 2.

The multi-channel signal $\widehat{\mathbf{x}}^M$, approximated by multi-channel Matching Pursuit, allows to estimate each mono-channel source signal s_i using the atoms of the decomposition that are allocated to it, in the following manner : assuming the mixing matrix $\mathbf{A}$ is known with unit columns $\|\mathbf{a}_i\|_2 = \sum_n a_{n,i}^2 = 1$, the atom k_M is attributed to the source of index :

$$\hat{i}_M = \arg\max_i |\langle \mathbf{c}_{k_M}, \mathbf{a}_i \rangle|.$$

This corresponds to partitioning the multi-channel coefficient space $\{\mathbf{c} = (c_n)_{1 \leq n \leq N} \in \mathbb{C}^N\}$ into I subsets corresponding to the columns $\mathbf{a}_i$ of $\mathbf{A}$ (I being the number of sources). The source s_i is reconstructed by :

$$\hat{s}_i = \sum_{M | \hat{i}_M = i} \langle \mathbf{c}_{k_M}, \mathbf{a}_i \rangle \phi_{k_M}. \tag{2}$$

We call this separation algorithm MPC1. Alternately, MPC2 is a variant consisting in attributing each atom to the N closest sources. This second selection, also used in Bofill & Zibulevsky's algorithm [8] in the stereophonic case ($N = 2$), corresponds to the minimization of the l_1 norm of the projection of the coefficients $\mathbf{c}_{k_M}$ on N directions of the mixing matrix :

$$\hat{J}_M = \arg\min_{J \subset [1,I]} \|\mathbf{A}_J^{-1} \mathbf{c}_{k_M}\|_1 \text{ , with } \mathbf{A}_J = [\mathbf{a}_i]_{i \in J} \tag{3}$$

3.2 Directional Multi-channel Dictionary

Combining the expression of the linear instantaneous mixtures $x_n = \sum_{i=1}^{I} a_{n,i} \cdot s_i$ and that of a candidate sparse decomposition $s_i = \sum_{k=1}^{K} c_{i,k}\phi_k$ of each source s_i on the mono-channel dictionary $\mathcal{D}$, we can write $x_n = \sum_{i,k} a_{n,i}\, c_{i,k}\, \phi_k$. This is translated in linear algebra as $\mathbf{x} = \mathbf{A}\mathbf{C}\mathbf{\Phi}^T$, with $\mathbf{\Phi}^T$ the matrix which rows are the mono-channel atoms ϕ_k, and $\mathbf{C} = \{c_{i,k}\}_{i,k}$ a matrix of sparse components. This decomposition can also be written $\mathbf{x} = \sum_{i,k} c_{i,k}\, \mathbf{a}_i\, \phi_k$, that is to say that $\mathbf{x}$ admits a sparse decomposition on the "directional" multi-channel dictionary constituted of the atoms $\mathbf{a}_i\, \phi_k = (a_{1,i}\, \phi_k, \ldots, a_{N,i}\, \phi_k)$. One can therefore get a decomposition of this type by applying MP on the latter dictionary. The inner products are then computed as $\langle \mathbf{R}^M, \mathbf{a}_i \phi_k \rangle = \mathbf{a}_i^T \mathbf{R}^M \phi_k^T$ and the source s_i is reconstructed by :

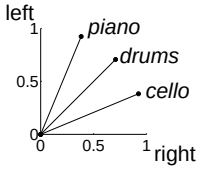
$$\hat{s}_i = \sum_k c_{i,k}\phi_k. \tag{4}$$

This new algorithm is called MPD and its theoretical properties have been studied in [11]. Using a directional dictionary is equivalent to applying multi-channel MP with the constraint that the components $\mathbf{c}_{k_M}$ of section 3.1 shall be proportional to a column $\mathbf{a}_i$ of $\mathbf{A}$.

4 Experiments

We compare the algorithms MPC1, MPC2 and MPD described previously to three reference algorithms. The experiments are performed on a stereophonic linear instantaneous mixture of three musical sources (a cello, some drums and a piano). The sampling frequency of the signals is 8kHz, and their length is 2.4s (19200 samples). The mixing matrix is the following :

$$\begin{bmatrix} \cos(\pi/8) & \cos(\pi/4) & \cos(3\pi/8) \\ \sin(\pi/8) & \sin(\pi/4) & \sin(3\pi/8) \end{bmatrix}$$



The energy of the drums, located in the middle, is about twice weaker than the energies of the piano and cello, which are quite similar.

We use the measures of separation performance proposed in [7], that allow to finely analyze the origin of the distortions between the estimated source and the original one. These measures, expressed in decibels, are based on the decomposition of an estimated source signal into parts due to original source, interferences and algorithmic artifacts. The relative ratios between the energies of these three parts define the Source to Distortion Ratio (SDR, global distortion), Source to Interference Ratio (SIR) and Source to Artifacts Ratio (SAR). For these three measures, of the same nature as the classical Signal to Noise Ratio, higher ratio mean better performances.

4.1 Reference algorithms

The performance of MPC1, MPC2 and MPD are compared to those of three reference algorithms : the best linear separator (BLS) [7], DUET [6] and the Bofill & Zibulevski's algorithm (BZ) [8].

The first one only consists in the application of a matrix $\mathbf{B}$ to the signal. $\mathbf{B}$ is such that the estimated sources $\hat{\mathbf{s}} = \mathbf{Bx}$ minimize the distortion due to the interferences [7]. If the sources are assumed to be mutually orthogonal, if the mixing matrix $\mathbf{A}$ is known, and if we denote $\mathbf{D}$ the diagonal matrix of the norms of the sources, then, with $\hat{\mathbf{A}} = \mathbf{AD}$, the matrix $\mathbf{B}$ is given by : $\mathbf{B} = \mathbf{D}\hat{\mathbf{A}}^H(\hat{\mathbf{A}}\hat{\mathbf{A}}^H)^{-1}$.

The algorithm DUET [6] applies a short-time Fourier transform (STFT) to each channel of the signal, then applies a mask that assumes only one source to be active for each time-frequency "box", and finally inverts the STFT to construct the estimated source.

The Bofill & Zibulevski's algorithm [8] relies on the same principle as DUET, the only difference being that each time-frequency box is attributed to the two nearest sources. This attribution is determined by an l_1 norm minimization (see Eq.3.)

In all the experiments, DUET and BZ are applied with a Hanning window of 4096 samples, with an overlap of 2048 samples (50% of the size of the window). Their performances strongly depend on the size of the window, and we have observed that a

greater, or more critically, smaller window size strongly decreases the performances in the studied cases. Therefore, the results shown below employ an *a posteriori* optimal window size. Note that in practice it might be hard to choose the optimal window size, since the performances can't be not known.

4.2 Different versions of MP algorithms

In this experiment, we study the influence of the number of iterations, of the composition of the dictionary, of the exploitation of the residual, and of a smoothing post-treatment using the MPD algorithm. Two dictionaries may be used for the decomposition :

- a "small" dictionary made of Gabor atoms of length $s = 4096$ with an overlap of half the length ($u = ns/2$, $n \in \mathbb{N}$). This corresponds to the STFT used by the DUET and BZ algorithms.
- a "large" dictionary made of Gabor atoms which length goes from $s = 64$ to 16384 (by powers of two). The overlap between two successive atoms is also 50% of the length of an atom.

Figure 1 represents the SDR, SIR and SAR of the "piano" source estimated by the different algorithms, against the number of iterations (the results are similar for the two other sources).

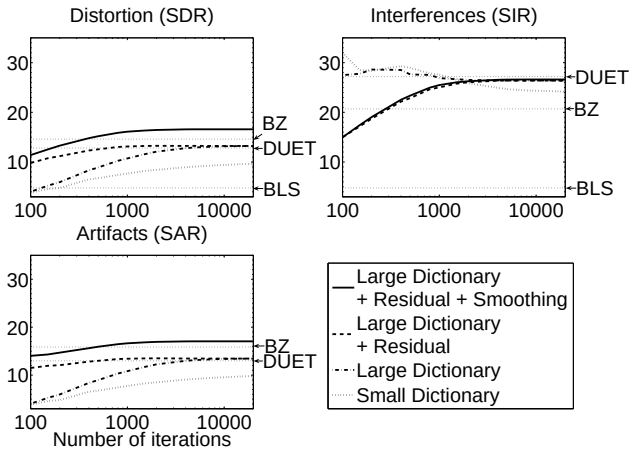


Figure 1. Distortions (dB), "piano" source estimated by MPD

Firstly, we can remark that for any number of iterations, using the large dictionary leads to a better separation than using the small dictionary. Indeed, in the case of the large dictionary, MP chooses the optimal window size automatically. The need to optimize *a priori* the window size is removed, contrarily to the BZ and DUET algorithms.

In addition, we can notice that the performance improvement is monotonic when the number of iterations increases. More precisely, artifacts, which dominate the distortion, are important when the sources are reconstructed with few atoms, and decrease

when more iterations are performed, thanks to the contribution of new atoms. After a sufficient number of iterations, MPD becomes better than DUET in terms of artifacts (SAR) and global distortion (SDR).

In order to compensate for the distortion due to the small number of atoms, the residual of the decomposition $\mathbf{R}^M$ can be separated using the linear separator $\mathbf{A}^H(\mathbf{A}\mathbf{A}^H)^{-1}\mathbf{R}^M$, and then added to the estimated sources. The separator we use assumes that all the residuals of the sources have the same energy. Asymptotically, the hypothesis is verified, the more energetic sources having their atoms selected in the first place. Adding the residual largely increases the global performance (SDR) for a small number of iterations. This effect decomposes into a strong diminution of artifacts and an increase of the interferences. For a larger number of iterations, the energy of the residual comes close to zero and the improvement brought on the SDR is less significant. The reduction of the artifacts allows to obtain better global performances than DUET with only a small number of iterations (less than 1000). Note that we could expect good improvement by using DUET or BZ instead of linear separator in this residual separation step.

Nevertheless, the performances of MPD plus separation of the residual are still worse that those of the BZ algorithm with an optimized window. Using the hypothesis that the smoothing introduced by the overlap of the windows of the STFT in the BZ algorithm plays a role in its good performance [12], we tried to smooth the sources estimated by MPD with the residual added. This smoothing consists in performing several estimations of the sources from shifted versions of the dictionary and producing the mean of these estimations. The amelioration brought by the smoothing, very clear for the artifacts (SAR improved by $\sim$ 3dB), but not systematic for the interferences (SIR), corresponds to better performances in SDR and SAR than DUET and BZ. For a completely fair comparison, the same smoothing should have been applied on DUET and BZ. Nevertheless, we can expect a lower improvement since this effect is already included in the methods.

For MPC1 and MPC2, changing the dictionary and adding the residual and the smoothing produce the same type of effects than for MPD. The iterative character of these algorithms implies a heavy computational cost, made tractable by a fast implementation of the algorithms [13].

4.3 What if the mixing matrix is imprecisely known ?

The following experiment evaluates the capacity of the different algorithms to maintain a good separation when the mixing matrix is no longer known, but only estimated. A voluntary imprecision is introduced by a rotation of the true matrix. The directions of the three sources are shifted by the same angle, which varies between $-\pi/16$ and $\pi/16$. The experiments are done with the "large" dictionary. They include the separation of the residual and the smoothing, and use 5000 iterations. The performances are given on Figure 2, depending on the perturbation angle, for the piano.

Evolution of the SAR – The studied methods keep an approximately constant level of artifacts for any angle of perturbation. MPD and BZ introduce the least artifacts, followed by MPC1 and MPC2 that present equivalent performances, and better than DUET. The levels of artifacts are intrinsic to the underlying models of each method.

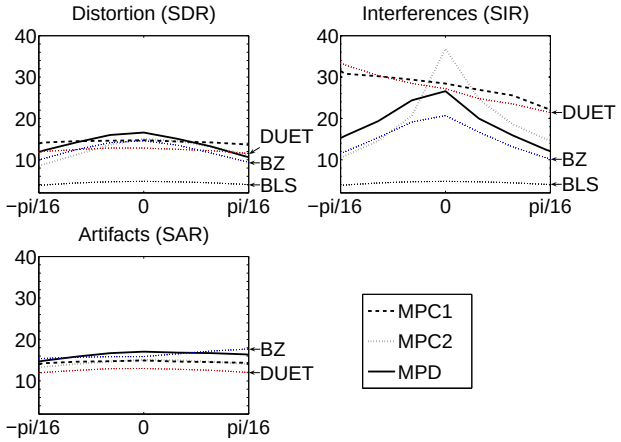


Figure 2. Distortions (dB), source "piano" depending on the perturbation angle

Evolution of the SIR – For the methods MPC1 and DUET, the time-frequency atoms are only attributed to one source. Therefore, these methods produce the least interferences and stay robust to a perturbation of the mixing matrix. In the case of MPC2, MPD and BZ, allotting time-frequency atoms to several sources introduces a larger sensibility to the perturbation on the mixing matrix. For a well-estimated mixing matrix, MPC2 produces the least interferences.

Evolution of the SDR – By definition, global distortion (SDR) is dominated by the minimum of SAR and SIR. For a well-estimated mixing matrix, the methods based on Matching Pursuit obtain better global performances than the reference methods. By decreasing order of performance, the methods are scaled as : MPD, MPC1, MPC2, BZ, DUET, and MSL. On the other hand, when a perturbation is introduced on the mixing matrix, the methods MPC1 and DUET (attribution to one direction) prove to be more robust than MPD (selection of the atoms by Matching Pursuit only on the estimated directions of the sources) and than the methods MPC2, BZ (attribution to two directions, that lead to a larger sensibility to interferences).

5 Conclusions

We have compared several methods for under-determined source separation by sparse decomposition, assuming that the mixing matrix is known. In the algorithms MPC1 and MPC2, the mixing matrix is used *a posteriori* to classify and gather the atoms resulting from the decomposition by Matching Pursuit. In the algorithm MPD, the knowledge of the mixing matrix is included *a priori* in the definition of the dictionary. The version of MPD with separation of the residual and the addition of the smoothing gives better performances, for global distortion and artifacts, than the methods DUET and BZ. When the mixing matrix is well estimated, MPD and MPC2 give the best results. On the other hand, MPC1 and DUET seem to be more robust to an error on the estimation of the mixing matrix.

The proposed formalism allows to perform separation in the case of underdetermined convolutive mixtures, provided that the mixing filters are known. In that case, the atoms of the multi-channel dictionary represent on each channel what is obtained at the sensor when each mono-channel atom is passed through the mixing filters. The algorithm is then just the application of Matching Pursuit on these normalized multi-channel atoms and the sources are reconstructed as in MPD. The related experiments are currently being developed.

Another perspective is to consider the joint estimation of the mixing matrix and the sources in the linear instantaneous case, or of the filters and the sources in the convolutive case. Alternately, we are investigating possible improvements of the sparse decomposition by learning dictionaries adapted to the mixture, notably directional multichannel dictionaries.

References

1. J.-F. Cardoso, "Blind signal separation: statistical principles," *Proc. IEEE, Special issue on blind identification and estimation*, vol. 9, no. 10, pp. 2009–2025, Oct. 1998.
2. O. Bermond and J.-F. Cardoso, "Mthodes de sparation de sources dans le cas sous-dtermin," in *Proc. GRETSI, Vannes, France*, 1999, pp. 749–752.
3. M. Zibulevsky and B. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Computations*, vol. 13, no. 4, pp. 863–882, 2001.
4. S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. on Signal Processing*, vol. 41, pp. 3397–3415, Dec. 1993.
5. R. Gribonval, "Sparse decomposition of stereo signals with matching pursuit and application to blind separation of more than two sources from a stereo mixture," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP'02), Orlando, Florida, USA, May 2002*, 2002.
6. O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.
7. R. Gribonval, L. Benaroya, E. Vincent, and C. F´evotte, "Proposals for performance measurement in source separation," in *Proc. 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, Apr. 2003, pp. 763–768, see "BSS EVAL Toolbox", `http://bass--db.gforge.inria.fr`.
8. P. Bofill and M. Zibulevsky, "Blind separation of more sources than mixtures using sparsity of their short-time fourier transform," in *Proc. ICA2000*, Helsinki, june 2000, pp. 87–92.
9. R. Gribonval, "Piecewise linear source separation," in *Proc. SPIE'03 – "Wavelets: Applications in Signal and Image Processing", San Diego, California, USA*, vol. 5207, August 2003, pp. 297–310.
10. B. D. Rao, S. Cotter, and K. Engan, "Diversity measure minimization based method for computing sparse solutions to linear inverse problems with multiple measurement vectors," in *Proceedings in Acoustics, Speech, and Signal Processing (ICASSP' 04)*, may 2004.
11. R. Gribonval and M. Nielsen, "Beyond sparsity : recovering structured representations by l1-minimization and greedy algorithms. – application to the analysis of sparse underdetermined ICA–," IRISA, Tech. Rep. 1684, Jan. 2005, http://www.irisa.fr/metiss/gribonval/.
12. S. Araki, S. Makino, H. Sawada, and R. Mukai, "Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask," in *Proceedings in Acoustics, Speech, and Signal Processing (ICASSP' 05)*, vol. 3, march 2005, pp. 81–84.
13. R. Gribonval and S. Krstulovi´c, "The Matching Pursuit ToolKit", see `http://mptk.gforge.inria.fr`.